Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences

Mukund V. Katti, Prabhakar K. Ranjekar, and Vidya S. Gupta

Plant Molecular Biology Unit, Division of Biochemical Sciences, National Chemical Laboratory, Pune, India

Complete chromosome/genome sequences available from humans, *Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana,* and *Saccharomyces cerevisiae* were analyzed for the occurrence of mono-, di-, tri-, and tetranucleotide repeats. In all of the genomes studied, dinucleotide repeat stretches tended to be longer than other repeats. Additionally, tetranucleotide repeats in humans and trinucleotide repeats in Drosophila also seemed to be longer. Although the trends for different repeats are similar between different chromosomes within a genome, the density of repeats may vary between different chromosomes of the same species. The abundance or rarity of various di- and trinucleotide repeats in different genomes cannot be explained by nucleotide composition of a sequence or potential of repeated motifs to form alternative DNA structures. This suggests that in addition to nucleotide composition of repeat motifs, characteristic DNA replication/repair/recombination machinery might play an important role in the genesis of repeats. Moreover, analysis of complete genome coding DNA sequences of Drosophila, *C. elegans,* and yeast indicated that expansions of codon repeats corresponding to small hydrophilic amino acids are tolerated more, while strong selection pressures probably eliminate codon repeats encoding hydrophobic and basic amino acids. The locations and sequences of all of the repeat loci detected in genome sequences and coding DNA sequences are available at http://www.ncl-india.org/ssr and could be useful for further studies.

Introduction

Simple sequence repeats (SSRs), or microsatellites, are the genetic loci where one or a few bases are tandemly repeated for varying numbers of times. Such repetitions occur primarily due to slipped-strand mispairing and subsequent error(s) during DNA replication, repair, or recombination (Levinson and Gutman 1987). These loci mutate by insertions or deletions of one or a few repeat units, and the mutation rates generally increase with an increase in the length of repeat tracks (Wierdl, Dominska, and Petes 1997). Microsatellite loci show extensive length polymorphism, and hence they are widely used in DNA fingerprinting and diversity studies. Moreover, since they are densely interspersed in eukaryotic genomes and can be easily assayed by PCR using unique flanking primers, they are considered ideal genetic markers for the construction of high-density linkage maps (Beckmann and Soller 1990; Morgante and Olivieri 1993).

Like any other regions of DNA, SSRs can also originate in coding regions, leading to the appearance of repetitive patterns in protein sequences. In protein sequence database studies, we have observed that tandem repeats are common in many proteins (Katti et al. 2000), and mechanisms involved in their genesis may contribute to the rapid evolution of proteins (Green and Wang 1994; Huntley and Golding 2000). During the past decade, several human neurodegenerative diseases have been found to be associated with dynamic mutations occurring at microsatellite loci within or near specific genes (Ashley and Warren 1995), leading to an increased interest in understanding the molecular mechanisms involved in the origin, evolution, and expansion/deletion of microsatellites.

Frequencies of various microsatellite sequences in different genomes have been estimated experimentally by hybridization techniques (e.g., Tautz and Renz 1984; Panaud, Chen, and McCouch 1995). However, this could not be done accurately using oligonucleotides like $(AT)_n$ and $(GC)_n$ that can self-complement. With the growth of sequence databases, several authors have reported an abundance of simple sequence repeats in different genomes (e.g., Hancock 1995; Jurka and Pethiyagoda 1995; Richard and Dujon 1996; Bachtrog et al. 1999; Kruglyak et al. 2000). In a recent survey, Toth, Gaspari, and Jurka (2000) examined the distribution of microsatellites in exonic, intronic, and intergenic regions of several eukaryotic taxa. Differential abundance of repeats in different genomes led them to suggest that strand-slippage theories alone are insufficient to explain characteristic microsatellite distributions.

Most of the previous studies on microsatellite distribution were based on DNA sequence databases in which coding or gene-rich regions were overrepresented. On the other hand, the availability of complete genome sequences now permits the determination of frequencies of SSRs at the whole-genome level. Such estimates should reflect the basal level of SSR dynamics within a species. The present paper details occurrences of SSRs in eukaryotic genomes that have been completely sequenced or for which complete chromosome sequences are available. Moreover, nonredundant complete genome-coding DNA sequences of Drosophila, *Caenorhabditis elegans,* and yeast have been analyzed to assess the extent of codon reiterations in protein-coding regions.

Materials and Methods

All the genome sequences were downloaded in FASTA format from ftp://ncbi.nlm.nih.gov/genbank/genomes/. The list of genome sequences and their

Key words: microsatellites, DNA strand slippage, codon repeats, genome sequences, database.

Address for correspondence and reprints: Vidya S. Gupta, Plant Molecular Biology Unit, Division of Biochemical Sciences, National Chemical Laboratory, Pune 411 008, India. E-mail: vidya@ems.ncl.res.in.

Mol. Biol. Evol. 18(7):1161–1167. 2001

Table 1
Frequencies of Repeat Loci per Million Base Pairs of Individual Chromosome Sequences in Different Eukaryotic Genomes

Chromosome/arm	Sequence Length (1,000,000 bp)	Frequency of Repeats ≥20 nt				Frequency of Repeats ≥40 nt			
		Mononucleotide Repeats	Dinucleotide Repeats	Trinucleotide Repeats	Tetranucleotide Repeats	Mononucleotide Repeats	Dinucleotide Repeats	Trinucleotide Repeats	Tetranucleotide Repeats
Human									
Hs-21	33.82	141.8	105.0	24.8	119.7	3.7	21.3	2.4	15.1
Hs-22	33.62	223.4	81.0	39.0	151.5	4.8	17.4	2.9	17.3
Drosophila									
Dm-X	21.95	157.0	215.1	135.8	96.8	0.8	9.5	7.3	4.2
Dm-2L	22.58	47.5	94.6	62.3	51.9	0.2	2.1	1.8	1.0
Dm-2R	21.07	45.4	102.7	79.0	57.4	0.3	3.3	2.9	1.7
Dm-3L	23.67	56.2	92.3	83.0	55.4	0.3	2.2	2.7	1.2
Dm-3R	27.86	53.8	104.9	85.0	58.0	0.3	3.5	2.5	1.5
Arabidopsis									
At-2	19.65	53.5	51.1	44.2	18.8	0.7	7.8	1.37	0.1
At-4	17.55	53.6	53.6	48.0	17.7	0.5	6.8	1.48	0.2
Caenorhabditis elegans									
Ce-I	14.75	37.5	34.8	28.8	21.2	0.1	4.7	0.61	0.6
Ce-II	16.62	30.4	22.4	25.8	25.3	0.1	3.1	0.60	0.4
Ce-III	11.60	30.3	30.9	31.8	19.4	0.0	3.7	0.43	0.3
Ce-IV	14.45	23.2	22.0	23.9	23.9	0.0	2.1	0.21	0.5
Ce-V	20.52	27.6	17.4	18.1	18.4	0.1	2.9	0.24	1.1
Ce-X	17.29	30.8	30.0	20.2	15.3	0.2	4.1	0.40	0.2
Yeast, all 16 chromosomes	12.07	44.2	31.7	50.0	12.3	1.8	2.4	4.89	0.3

lengths are shown in table 1. The human chromosome 21 (Hattori et al. 2000) and chromosome 22 (Dunham et al. 1999) sequences were obtained as ensembles of 5 and 12 contig sequences, respectively. Individual chromosome sequences of *Saccharomyces cerevisiae* (Goffeau et al. 1996), *C. elegans* (*C. elegans* Sequencing Consortium 1998), and *Arabidopsis thaliana* chromosomes II (Lin et al. 1999) and IV (Mayer et al. 1999) were available as single contiguous strings. The *C. elegans* chromosome sequences had a few unsequenced gaps represented as stretches of ''N'' in the sequences, and the lengths shown in table 1 are corrected by removing such gaps. Most of the *Drosophila melanogaster* genome has been sequenced by whole-genome shotgun sequencing (Adams et al. 2000), and sequences have been made available as a collection of scaffolds. Only the genomic scaffolds mapped on chromosomes X, 2, and 3 were selected and obtained using GenBank's Batch Entrez facility. Accession numbers or links to all of the sequences used in this study are available at http://www.ncl-india.org/ssr.

All of the genome sequences were scanned for various SSRs using computer programs written in C. A simple sliding-window technique was used for detection of tandem repeats. Briefly, consider a DNA sequence as a string, $B_1B_2B_3B_4B_5 \ldots B_i \ldots B_{n-1}B_n$. To detect a tandem repeat of size ($k = 1$–4) at position i, the window $B_i \ldots B_{i+k-1}$ was compared with subsequent windows starting at positions B_{i+k}, B_{i+2k}, B_{i+3k}, B_{i+4k}, ... A repeat was detected and extended further when a certain minimum number of units (20, 10, 7, or 5 for mono-, di-, tri-, or tetranucleotide repeats, respectively) were repeated tandemly. Repeats were searched allowing a maximum of one mismatch for every 10 nt. While scanning for di-, tri-, and tetranucleotide repeats, combinations involving runs of same nucleotide were not considered. Similarly, for tetranucleotide repeats, combinations representing perfect dinucleotide repeats were ignored. The significance of the difference in density of repeats between different chromosomes of the same species was determined using a *t*-test. Frequency distributions of repeats along one million-bp contiguous segments of a chromosome were used for calculation of variance for the *t*-test. However, the significance could be tested only for human, Arabidopsis, and *C. elegans* chromosome sequences, for which long contiguous chromosome sequences were available.

A polyA repeat is same as a polyT repeat on a complementary strand. Similarly, $(AC)_n$ is equivalent to $(CA)_n$, $(TG)_n$, and $(GT)_n$, while $(AGC)_n$ is equivalent to $(GCA)_n$, $(CAG)_n$, $(CTG)_n$, $(TGC)_n$, and $(GCT)_n$ in different reading frames or on a complementary strand. Thus, 2 unique classes are possible for mononucleotide repeats, whereas 4 classes are possible for dinucleotide, 10 for trinucleotide, and 33 for tetranucleotide repeats (Jurka and Pethiyagoda 1995). We determined individual repeat frequencies for all of these classes.

Complete genome coding DNA sequences of all predicted peptides of Drosophila, *C. elegans*, and yeast were obtained from the Berkeley Drosophila Genome Project (http://www.fruitfly.org), the Sanger Centre's Wormpep Database (http://www.sanger.ac.uk/Projects/C_elegans/wormpep), and the *Saccharomyces* Genome Database (http://genome-www.stanford.edu/Saccharomyces), respectively. A codon repeat was considered

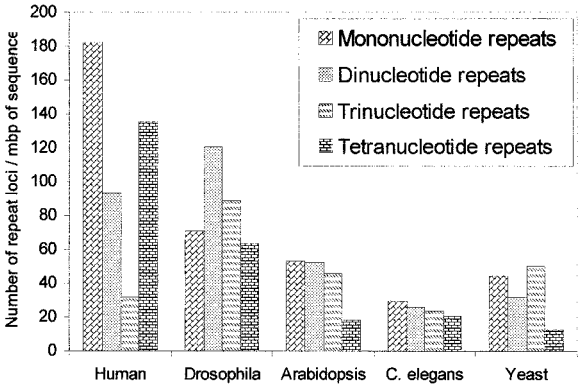

FIG. 1.—Frequencies of repeat loci per million base pairs of chromosome sequences in different genomes.

only when it was tandemly repeated for a minimum of seven times allowing one mismatch for every 10 nt.

Results and Discussion

While searching a sequence for simple sequence repeats, definition of the minimum number of repeats and mismatch considerations are important empirical criteria. For detection of various repeats in genome sequences, we selected minimum repeating units such that a repeat spanned a minimum of 20 nt. Although previous studies have used threshold repeat lengths of 10–12 nt, any preference(s) in genesis of repeats or variations in mutation rates are likely to be clearer at longer threshold lengths. Additionally, longer repeats, being more unstable, have implications in genome organization, genetic variation, protein evolution, and disease on a relatively shorter evolutionary timescale. Simple sequences can be pure tandem repeats or contain interruptions due to accumulation of point mutations or have scrambled arrangements of repetitive motifs (Tautz, Trick, and Dover 1986). However, most of the previous studies have considered only perfect repeats, without allowing any mismatch. We observed that several long repeats contained one or a few base substitutions. Hence, if only perfect repeats are considered, such loci are likely to be counted as two or more separate repeats of shorter lengths. Therefore, rather than considering only perfect repeats, we allowed one mismatch for every 10 nt. Although the appearance of mismatches in repeats can reduce the chances of slippage-mediated expansions/deletions (Petes, Greenwell, and Dominska 1997), such loci might represent previous occurrences of perfect repeats. Moreover, interruptions in a repeat track may be only a transitional state and could be removed by DNA replication slippage or reverse mutations (Harr, Zangerl, and Schlotterer 2000).

Analysis of complete chromosome/genome sequences of humans, Drosophila, Arabidopsis, *C. elegans*, and yeast for occurrences of various microsatellites (table 1 and fig. 1) revealed that compared with other genomes, human chromosomes 21 and 22 are rich in mono- and tetranucleotide repeats. On the other hand, the Drosophila chromosomes have higher frequencies of di- and trinucleotide repeats. Surprisingly, the *C. elegans* genome contains less SSRs per million base pairs of sequence compared to that in the yeast genome. Moreover, the frequency of trinucleotide repeats in yeast is higher than that observed in human chromosomes 21 and 22.

In all of the genomes, among mononucleotide repeats, polyA/polyT repeats were predominant, while polyC/polyG repeats were rare. Tetranucleotide repeats were very frequent in human chromosomes, and the most common among them were $(AAAT)_n$, $(AAAG)_n$, $(AAAC)_n$, $(ATAG)_n$, $(AAGG)_n$, $(ATGG)_n$, and $(AGGG)_n$. The Drosophila chromosomes also contained a large number of tetranucleotide repeats, of which $(ATAC)_n$, $(AAAT)_n$, $(AAAC)_n$, $(AGTC)_n$, and $(AACC)_n$ were more frequent. Overall, tetranucleotide repeats of type $(AAAN)_n$ seemed to be more common than other combinations.

The length distributions of all SSRs indicated that the frequency of repeats decreases exponentially with repeat length (data not shown). This may be because longer repeats have higher mutation rates and hence are more unstable (Wierdl, Dominska, and Petes 1997; Kruglyak et al. 1998). The paucity of longer microsatellites could also be due to their downward mutation bias and short persistence time (Harr and Schlotterer 2000). Recent studies have shown that compared with expansion mutation events, contraction mutations occur more frequently with increases in allele size (Xu et al. 2000), and long alleles tend to mutate to shorter lengths, thus preventing their infinite growth (Ellegren 2000).

Among the repeats longer than ~40 nt, the dinucleotide repeats were more frequent, whereas mononucleotide repeats seemed to be less common (table 1). Large numbers of tetranucleotide repeats in human chromosomes and trinucleotide repeats in Drosophila were also longer than ~40 nt. Slippage rates have been estimated to be highest in dinucleotide repeats, followed by tri- and tetranucleotide repeats (Chakraborty et al. 1997; Kruglyak et al. 1998; Schug et al. 1998). Probably, shorter repeating units allow more possible slippage events per unit length of DNA and hence are likely to be more unstable. However, shorter lengths of mononucleotide repeats in all genome sequences and an abundance of tetranucleotide repeats in human sequences suggest the involvement of additional mechanisms.

Our study shows that compared with human chromosome 21, chromosome 22 has significantly higher frequencies of mono-, tri-, and tetranucleotide repeats but lower frequencies of dinucleotide repeats (*t*-test: $t = 5.60$ for mononucleotide repeats, $t = 3.42$ for dinucleotide repeats, $t = 4.59$ for trinucleotide repeats, and $t = 3.94$ for tetranucleotide repeats; $P < 0.01$ in all the cases). In *C. elegans*, among a total of 60 chromosome pair/repeat type combinations, 15 combinations showed significant differences in density of repeats (at $P < 0.05$). On the other hand, the densities of repeats in Arabidopsis chromosomes 2 and 4 were similar. For *Drosophila*, the sex chromosome (X) contained ~1.5–3 times as many repeats per million base pairs of sequence as autosomes (chromosomes 2 and 3) (significance not calculated). Such differences for dinucleotide repeats in the

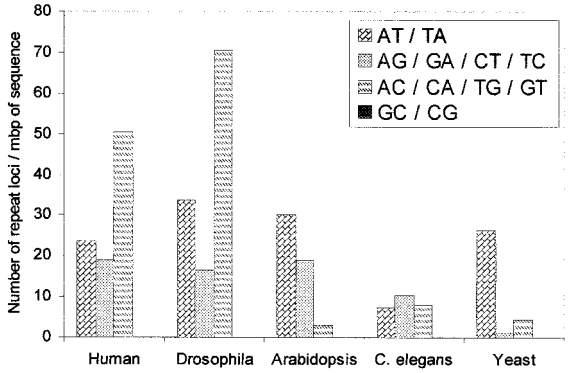

FIG. 2.—Frequencies of different dinucleotide repeats per million base pairs of chromosome sequences in different genomes.

Drosophila sex chromosome and autosomes have been reported (Pardue et al. 1987; Bachtrog et al. 1999). Thus, although the trends for different repeat classes are similar between chromosomes within a genome, the density of repeats may vary between different chromosomes of the same species. This can be expected, since different chromosomes in a genome can have different organizations of genes, euchromatin, and heterochromatin.

Relative Frequencies of Various Di- and Trinucleotide Repeats

All dinucleotide repeat combinations excluding homomeric dinucleotides can be grouped into four unique classes, namely, $(AT)_n$, $(AG)_n$, $(AC)_n$, and $(GC)_n$. It is evident that in human and Drosophila chromosomes, AC dinucleotide repeats are more frequent, followed by AT and AG repeats (fig. 2). In contrast, Arabidopsis chromosomes contain more AT repeats, followed by AG repeats. However, in the yeast genome, AT repeats seem to be predominant compared with other dinucleotide repeats. Interestingly, GC dinucleotide repeats are extremely rare in all of the genomes studied. Lower frequencies of CpG dinucleotides in vertebrate genomes has been attributed to methylation of cytosine, which, in turn, increases its chances of mutation to thymine by deamination (Schorderet and Gartler 1992). However, CpG suppression by this mechanism cannot explain the rarity of $(CG)_n$ dinucleotide repeats in yeast, *C. elegans*, and Drosophila, since they do not show cytosine methylation.

Among 10 unique trinucleotide repeat classes, human chromosomes 21 and 22 contain more AAT and AAC repeats (fig. 3). Compared with other genomes, Drosophila chromosomes have the highest frequency of trinucleotide repeats, and among them, AGC repeats are predominant, followed by AAC repeats. The Arabidopsis and *C. elegans* chromosomes have comparatively higher frequencies of AAG trinucleotide repeats. In contrast, the yeast genome contains more AAT, AAG, AAC, ATG, and AGC repeats. It should be noted that frequencies of trinucleotide repeats in the chromosome sequences also include those occurring in the coding regions and could be partially limited by selection at the protein level.

Short protomicrosatellites are probably generated by random mutations and then expand by DNA-slippage-mediated events. Therefore, the base composition of a sequence that provides seeds for evolution of repeats is expected to influence microsatellite density (Bachtrog et al. 1999; Kruglyak et al. 2000). We tested this assumption first by an XY scatter plot representation of percentages of di- and trinucleotide composition of a sequence and frequencies of corresponding repeats in individual chromosomes. It was observed that differences in frequencies of various repeat classes were large and could not be attributed to differences in nucleotide composition of a sequence (data not shown).

DNA strand slippage can occur during transient dissociation and reannealing in the repeat region, and this could be a deceptive event for DNA processing machinery leading to expansions or deletions in the repeat tracks. It has been suggested that if the nucleotides on the single strand are self-complementary, they can base-pair to form loops or hairpins and stabilize strand slip-

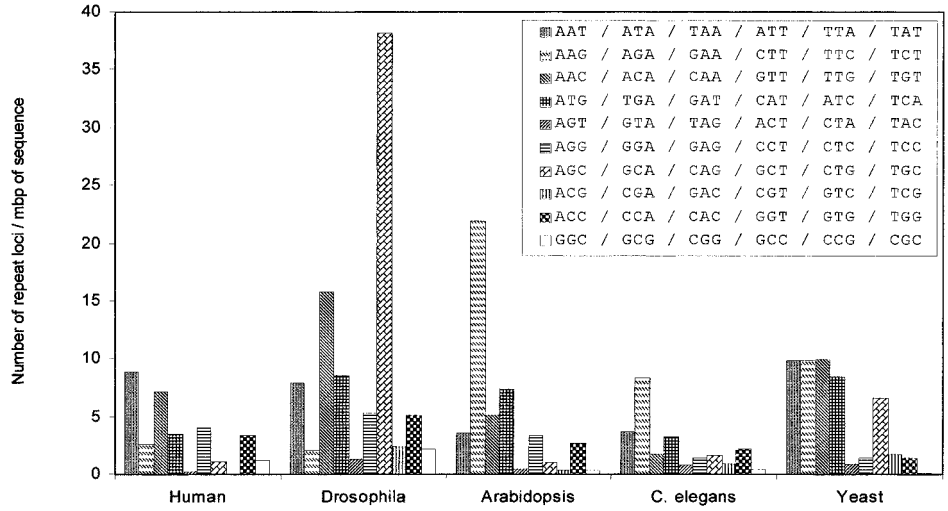

FIG. 3.—Frequencies of different trinucleotide repeats per million base pairs of chromosome sequences in different genomes.

Table 2
Total Occurrences of Codon Repeats in Complete Genome Coding DNA Sequence Sets of Drosophila, *Caenorhabditis elegans,* and Yeast

Codons	Encoded Amino Acid Residue	Drosophila		*C. elegans*		Yeast	
		Codon Repeated ≥7 Times	Codon Repeated ≥14 Times	Codon Repeated ≥7 Times	Codon Repeated ≥14 Times	Codon Repeated ≥7 Times	Codon Repeated ≥14 Times
GGA/GGG/GGC/GGT	Glycine	141	2	51	0	4	0
GCA/GCG/GCC/GCT	Alanine	274	14	49	0	13	0
GTA/GTG/GTC/GTT	Valine	4	0	7	0	3	0
CTA/CTG/CTC/CTTTA/TTG	Leucine	12	0	8	1	3	1
ATA/ATC/ATT	Isoleucine	10	0	5	0	0	0
TGC/TGT	Cysteine	3	0	1	0	1	0
ATG	Methionine	4	0	0	0	0	0
TAC/TAT	Tyrosine	2	1	7	0	2	0
TTC/TTT	Phenylalanine	4	0	9	0	10	1
TGG	Tryptophan	0	0	0	0	0	0
CCA/CCG/CCC/CCT	Proline	54	0	103	0	7	0
TCA/TCT/TCC/TCTAGC/AGT	Serine	250	9	29	0	34	2
ACA/ACG/ACC/ACT	Threonine	119	3	32	0	4	0
AAC/AAT	Asparagine	175	10	25	1	79	16
CAA/CAG	Glutamine	1,555	107	130	0	122	7
GAC/GAT	Aspartic acid	79	0	108	2	81	10
GAA/GAG	Glutamic acid	166	6	98	0	81	5
AAA/AAG	Lysine	47	0	78	0	22	0
CGA/CGG/CGC/CGTAGA/AGG	Arginine	2	0	9	2	4	1
CAC/CAT	Histidine	92	0	24	0	13	0
Total occurrences of repeats		2,993	152	773	6	483	43
Total coding sequences analyzed		14,080	14,080	19,209	19,209	6,283	6,283

page (Gacy et al. 1995; Moore et al. 1999). If these mechanisms favor repeat expansions/deletions, repeats with higher hairpin propensities like $(CTG)_n$ and $(CCG)_n$ (Gacy et al. 1995; Mitas et al. 1995) or self-complementary repeats like $(AT)_n$ and $(GC)_n$ are likely to be more abundant. However, relative frequencies of various di- and trinucleotide repeat classes within and between different genomes do not seem to support such an association. For example, trinucleotide repeats of the AGC class (representing CAG/CTG repeats) are predominant in Drosophila, whereas in humans, Arabidopsis, and *C. elegans* genome sequences, they are less frequent. In contrast, human chromosomes 21 and 22 contain more of AAT and AAC trinucleotide repeats, although their relative hairpin propensity is low (Gacy et al. 1995; Mitas et al. 1995). Similarly, trinucleotide repeats of the AAG class that can adopt triple-helical structures (Pearson and Sinden 1998) are comparatively more numerous in Arabidopsis, *C. elegans,* and yeast and less numerous in human and Drosophila sequences. This suggests that in addition to alternative DNA structures formed by repeat motifs, species-specific cellular factors interacting with them are likely to play an important role in the genesis of repeats (Toth, Gaspari, and Jurka 2000).

Codon Repetitions in Complete Genome Coding DNA Sequences

Among all SSRs, slippage-mediated expansions/deletions of only trinucleotide repeats or multiples thereof can be tolerated in coding regions, since they do not disturb the reading frame. Coding DNA sequences of all the predicted peptides of Drosophila, *C. elegans,* and yeast genomes were analyzed for the occurrence of the same codon (trinucleotide) consecutively repeated seven or more times (table 2). It is evident that codon repetitions are far more frequent in Drosophila than in *C. elegans*, which in fact has more predicted proteins than Drosophila. This is to be expected, since the frequency of microsatellites is very low in *C. elegans* (fig. 1). In Drosophila coding sequences, CAG codon (encoding glutamine) repetitions are predominant, followed by AGC (serine), GAG (glutamic acid), GCA (alanine), and AAC (asparagine) repeats. On the other hand, in *C. elegans* coding sequences, GAT (aspartic acid), CCA (proline), CAA (glutamine), GAA (glutamic acid), and AAG (lysine) codon repeats are comparatively more frequent, although very few of them are repeated 14 or more times. In yeast open reading frames (ORFs), GAA (glutamic acid), CAA (glutamine), GAT (aspartic acid), AAT (asparagine), and CAG (glutamine) codon repeats are more numerous. Such trends for triplet repeats in yeast ORFs have also been reported previously and are thought to reflect functional selection acting on amino acid reiterations in the encoded proteins (Alba, Santibanez-Koref, and Hancock 1999).

The correlation coefficient between frequencies of various trinucleotide repeat classes in coding sequences and in noncoding sequences (frequency in total genome sequences minus frequency in total coding sequences) was found to be significant in Drosophila ($r = 0.84$, $P < 0.01$) but insignificant in *C. elegans* ($r = 0.53$) and yeast ($r = 0.37$). It was also noted that within a trinucleotide repeat class, frequencies of different codon repeats vary considerably depending on the type of encoded amino acid. Perhaps the most interesting observation in our study is that expansions of codons corre-

sponding to small/hydrophilic amino acids are more tolerated than are hydrophobic amino acids, and this is particularly evident for codons repeated 14 or more times. Therefore, while nucleotide composition might play an important role in the genesis of repeats, in the coding sequences, its effect on the structure and function of the encoded proteins would be a major selective force. For example, at the DNA level, physical and chemical properties of $(AGC)_n$, $(GCA)_n$, $(CAG)_n$, $(CTG)_n$, $(TGC)_n$, and $(GCT)_n$ repeats are the same, and their frequencies can be expected to be comparable. However, in the Drosophila coding DNA sequence set, there are 204 occurrences of AGC (serine), 175 of GCA (alanine), 1,480 of CAG (glutamine), 36 of GCT (alanine), 11 of CTG (leucine), and 3 of TGC (cysteine) codon repeats (codons reiterated seven or more times).

The trends observed for codon repeats in complete genome coding DNA sequences are consistent with our previous study of a protein sequence database, where we observed that single amino acid repeat stretches of small/hydrophilic amino acids were more frequent in proteins (Katti et al. 2000). This might perhaps explain why the majority of the repeat-associated diseases are due to expansions of CAG repeats in specific genes. Since glutamine repeats are tolerated more in proteins, the initial small $(CAG)_n$ expansions in coding regions are likely to have enough survival value to remain in a population. However, as their instability increases with increasing length, their effect on protein structure and function could be deleterious beyond a certain limit, leading to the protein malfunctioning (Perutz 1999). On the other hand, initial small expansions of hydrophobic and basic amino acid residues could be lethal and hence would be eliminated from the population as soon as they appeared. The availability of a complete coding DNA sequence set of the human genome will enable us to test this hypothesis.

Conclusions

Analysis of SSRs in genome sequences gives a snapshot of in vivo accumulated repeats. Overall, the trends observed for various repeat classes in genome sequences are in agreement with previous reports (e.g., Richard and Dujon 1996; Bachtrog et al. 1999; Kruglyak et al. 2000; Toth, Gaspari, and Jurka 2000). However, with the availability of complete genome/chromosome sequences, we have begun to understand the extent to which repeats are generated in a genome. Differential distributions of various repeats observed in different genome sequences suggest that apart from the nucleotide composition of repeats, the characteristic DNA replication/repair/recombination machinery might have an important role in the evolution of SSRs. In addition, their occurrence in coding regions seems to be limited by nonperturbation of the reading frame and tolerance of expanding amino acid repeat stretches in the encoded proteins. These observations have implications for our efforts to understand the instability of disease-associated repeats.

The locations and sequences of all of the microsatellite loci reported in this study are available at http://www.ncl-india.org/ssr. This information could be useful for the selection of a wide range of microsatellite loci for studying their location and sequence-dependent evolution. They can also be used as markers for the fine analysis of recombination events along individual chromosomes. Availability of data on microsatellite content of complete chromosome sequences should also facilitate comprehensive studies on the direct role of microsatellites in genome organization, recombination, gene regulation, quantitative genetic variation, and evolution of genes.

Acknowledgments

M.V.K. acknowledges the Council of Scientific and Industrial Research (CSIR), New Delhi, India, for the award of a Senior Research Fellowship. We thank Prof. M. V. Hegde and Dr. R. Sami-Subbu for discussions, and Dr. V. Shankar for critical reading of the manuscript. We also thank the anonymous reviewers of the manuscript for many useful suggestions.

LITERATURE CITED

ADAMS, M. D., S. E. CELNIKER, R. A. HOLT et al. (195 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185–2195.

ALBA, M. M., M. F. SANTIBANEZ-KOREF, and J. M. HANCOCK. 1999. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J. Mol. Evol. **49**:789–797.

ASHLEY, C. T., and S. T. WARREN. 1995. Trinucleotide repeat expansion and human disease. Annu. Rev. Genet. **29**:703–728.

BACHTROG, D., S. WEISS, B. ZANGERL, G. BREM, and C. SCHLOTTERER. 1999. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. Mol. Biol. Evol. **16**:602–610.

BECKMANN, J. S., and M. SOLLER. 1990. Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. Biotechnology **8**:930–932.

C. ELEGANS SEQUENCING CONSORTIUM. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**:2012–2018.

CHAKRABORTY, R., M. KIMMEL, D. N. STIVERS, L. J. DAVISON, and R. DEKA. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94**:1041–1046.

DUNHAM, I., N. SHIMIZU, B. A. ROE et al. (217 co-authors). 1999. The DNA sequence of human chromosome 22. Nature **402**:489–495.

ELLEGREN, H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24**:400–402.

GACY, A. M., G. GOELLNER, N. JURANIC, S. MACURA, and C. T. MCMURRAY. 1995. Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. Cell **81**:533–540.

GOFFEAU, A., B. G. BARRELL, H. BUSSEY et al. (16 co-authors). 1996. Life with 6000 genes. Science **274**:546–567.

GREEN, H., and N. WANG. 1994. Codon reiterations and the evolution of proteins. Proc. Natl. Acad. Sci. USA **91**:4298–4302.

HANCOCK, J. M. 1995. The contribution of slippage-like processes to genome evolution. J. Mol. Evol. **41**:1038–1047.

HARR, B., and C. SCHLOTTERER. 2000. Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics **155**:1213–1220.

HARR, B., B. ZANGERL, and C. SCHLOTTERER. 2000. Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from Drosophila. Mol. Biol. Evol. **17**:1001–1009.

HATTORI, M., A. FUJIYAMA, T. D. TAYLOR et al. (63 co-authors). 2000. The DNA sequence of human chromosome 21. Nature **405**:311–319.

HUNTLEY, M., and G. B. GOLDING. 2000. Evolution of simple sequence in proteins. J. Mol. Evol. **51**:131–140.

JURKA, J., and C. PETHIYAGODA. 1995. Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. **40**:120–126.

KATTI, M. V., R. SAMI-SUBBU, P. K. RANJEKAR, and V. S. GUPTA. 2000. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. Protein Sci. **9**:1203–1209.

KRUGLYAK, S., R. T. DURRETT, M. D. SCHUG, and C. F. AQUADRO. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95**: 10774–10778.

———. 2000. Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Mol. Biol. Evol. **17**: 1210–1219.

LEVINSON, G., and G. A. GUTMAN. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4**:203–221.

LIN, X., S. KAUL, S. ROUNSLEY et al. (37 co-authors). 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana.* Nature **402**:761–768.

MAYER, K., C. SCHULLER, R. WAMBUTT et al. (230 co-authors). 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana.* Nature **402**:769–777.

MITAS, M., A. YU, J. DILL, T. J. KAMP, E. J. CHAMBERS, and I. S. HAWORTH. 1995. Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: $(CTG)_{15}$. Nucleic Acids Res. **23**:1050–1059.

MOORE, H., P. W. GREENWELL, C. P. LIU, N. ARNHEIM, and T. D. PETES. 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. Proc. Natl. Acad. Sci. USA **96**:1504–1509.

MORGANTE, M., and A. M. OLIVIERI. 1993. PCR-amplified microsatellites as markers in plant genetics. Plant J. **3**:175–182.

PANAUD, O., X. CHEN, and S. R. MCCOUCH. 1995. Frequency of microsatellite sequences in rice (*Oryza sativa* L.). Genome **38**:1170–1176.

PARDUE, M. L., K. LOWENHAUPT, A. RICH, and A. NORDHEIM. 1987. $(dC\text{-}dA)_n$. $(dG\text{-}dT)_n$ sequences have evolutionarily conserved chromosomal locations in *Drosophila* with implications for roles in chromosome structure and function. EMBO J. **6**:1781–1789.

PEARSON, C. E., and R. R. SINDEN. 1998. Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. Curr. Opin. Struct. Biol. **8**:321–330.

PERUTZ, M. F. 1999. Glutamine repeats and neurodegenerative diseases: molecular aspects. Trends Biochem. Sci. **24**:58–63.

PETES, T. D., P. W. GREENWELL, and M. DOMINSKA. 1997. Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae.* Genetics **146**:491–498.

RICHARD, G. F., and B. DUJON. 1996. Distribution and variability of trinucleotide repeats in the genome of the yeast *Saccharomyces cerevisiae.* Gene **174**:165–174.

SCHORDERET, D. F., and S. M. GARTLER. 1992. Analysis of CpG suppression in methylated and nonmethylated species. Proc. Natl. Acad. Sci. USA **89**:957–961.

SCHUG, M. D., C. M. HUTTER, K. A. WETTERSTRAND, M. S. GAUDETTE, T. F. MACKAY, and C. F. AQUADRO. 1998. The mutation rates at di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. Mol. Biol. Evol. **15**:1751–1760.

TAUTZ, D., and M. RENZ. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. **12**:4127–4138.

TAUTZ, D., M. TRICK, and G. A. DOVER. 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature **322**:652–656.

TOTH, G., Z. GASPARI, and J. JURKA. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. **10**:967–981.

WIERDL, M., M. DOMINSKA, and T. D. PETES. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146**:769–779.

XU, X., M. PENG, Z. FANG, and X. XU. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. **24**:396–399.

DIETHARD TAUTZ, reviewing editor

Accepted March 7, 2001